AutoML for End-to-End Clustering

1. General Info

Project Title: AutoML for End-to-End Clustering

Contact Person: Azade Farshad, Yousef Yeganeh

Contact Email: azade.farshad@tum.de, y.yeganeh@tum.de

2. Project Abstract

The goal of this project is to use AutoML techniques[1] for end-to-end clustering of data using ANNs. AutoML techniques are methods such as meta-learning (learning to learn) and neural architecture search which aim to automate the learning process using meta-features. We use GMVAE[2] as the clustering method and employ a RL-based approach for finding the optimum number of clusters[3] to adapt the GMVAE network layers for end-to-end clustering. To evaluate the method, synthetic data would be used in the first stage and then the method will be evaluated on medical imaging datasets such as HAM10k or any dataset of choice.

3. Background and Motivation

Recent advances in deep learning have moved toward using less supervision. Clustering is one of the important fundamental approaches in unsupervised learning for classification of data. End-to-End clustering methods [4] have been investigated in some works, but most of them are dependent on a predefined value for the number of clusters. A previous work [5] has focused on determining the optimum number of clusters using Reinforcement Learning. This approach could be used to adapt the layers of the GMVAE network to the predicted number of clusters for the task of end-to-end clustering using ANNs.

4. Technical Prerequisites

- Good background in statistics
- Good background in machine learning, deep learning
- Good skills in Python
- Good skills in PyTorch

5. Benefits:

- Possible novelty of the research
- Possible publication

6. Students' Tasks Description

Students' tasks would be the following:

Groups 1 & 2:

- Understanding the underlying methods
- Implementing and adapting GMVAE for data clustering
- Choosing the appropriate AutoML techniques for adapting the network layers to the predicted number of clusters
- Reimplementing GMVAE with adaptive layers

- Running the evaluation metrics on toy / synthetic datasets
- Running the evaluation metrics on a medical imaging dataset
- Testing and documentation.

7. Work-packages and Time-plan:

	Description	#Students	From	To
WP1	Familiarize with the literature.	4	22.04	29.04
WP2	Familiarize with the required frameworks. Come up with a detailed time-plan (gantt)	4	29.04	06.05
WP3	Implementing and adapting the GMVAE for clustering + Implementing the required evaluation measures	4	06.05	13.05
WP4	Implementing the adaptive layers for GMVAE	4	13.05	27.05
WP5	Evaluation of the implemented method	4	27.05	03.06
WP6	Comparison to related work + Preparing midterm presentation	4	03.06	10.06
M1	Intermediate Presentation II	4	10.06.2021	
WP7	Familiarize with clinical data, data pre-processing	4	10.06	17.06
WP8	Implement and Evaluate WP3 & WP6 on medical data	4	17.06	01.06
WP9	Testing and Documentation	4	01.07	15.07
M2	Final Presentation	4	15.07.2021	

References

1. Hutter, F., Kotthoff, L. and Vanschoren, J., 2019. *Automated machine learning: methods, systems, challenges* (p. 219). Springer Nature.
2. Dilokthanakul, N., Mediano, P. A., Garnelo, M., Lee, M. C., Salimbeni, H., Arulkumaran, K., & Shanahan, M. (2016). Deep unsupervised clustering with gaussian mixture variational autoencoders. arXiv preprint arXiv:1611.02648.
3. Garg, V., & Kalai, A. T. (2018). Supervising unsupervised learning. Advances in Neural Information Processing Systems, 31, 4991-5001.
4. Caron, M., Bojanowski, P., Joulin, A., & Douze, M. (2018). Deep clustering for unsupervised learning of visual features. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 132-149).
5. Meta-k: Towards Self-supervised Prediction Of Number Of Clusters